OXFORD GENETICS

Using Machine Learning Methods for Identification of lncRNA from Microarray Sepsis Gene Expression Datasets

Joseph Luper Tsenum[1], Milad Rad[2] and Rishikesan Kamaleswaran[3]

[1]School of Biological Sciences, Georgia Institute of Technology
[2,3]School of Biomedical Engineering, Georgia Institute of Technology
[2]School of Medicine, Emory University

*Corresponding Author: Joseph Luper Tsenum; School of Biological Sciences, Georgia Institute of Technology; jtsenum3@gatech.edu.

Abstract

Sepsis is a disease caused by dysregulated reaction to infection. Early diagnosis of sepsis is important to its treatment. Increasing evidence have demonstrated the roles of lncRNAs as potential biomarkers of many human diseases, including sepsis. Identifying lncRNAs as biomarkers of different diseases is a difficult task. The choice of feature selection methods and balancing small data sets is challenging. In recent years, machine learning methods have been successfully applied on microarray data for different predictive genomic studies. In the present study, we trained binary classification models that predicted lncRNAs as potential biomarkers in sepsis. Among our models, XGBoost performed best at AUC of 0.97. Based on XGBoost's predictive power, We identified 10 (ten) potential biomarkers for sepsis; namely WDFY3-AS2, NFE4, WDR11-AS1, TP53TG1, ZNF790-AS1, ANKRD10-IT1, LINC01352, CHRM3-AS2, LINC00528, FAM13A-AS1.

Keywords: Sepsis; long non-coding RNAs; Machine learning; Biomarkers

Introduction

Sepsis as defined by "The third international consensus is a life-threatening organ dysfunction that occurs due to a dysregulated host response to infection". In recent years, there has been increasing cases of sepsis ICU hospitalization. Septic shock on the other hand is a subtype of sepsis (Singer et al., 2016). Recent studies have demonstrated the role of lncRNAs, microRNAs (miRNAs) and circular RNAs as biomarkers of different human diseases (Hashemian et al., 2020). The poor prognosis of sepsis in the peripheral blood cells have been linked to upregulated lncRNA colorectal neoplasia differentially expressed (CRNDE) expression levels (Huang et al., 2019). Adelman et al (2020) reported lncRNA metastasis-associated lung adenocarcinoma transcript (MALAT) 1 promoted the expression of TNF- in LPS-treated cardiomyocytes. The role of lncRNAs as biomarkers of sepsis is not fully understood. There is need for more in silico, in vivo and in vitro experiments to identify and verify lncRNAs as biomarkers of sepsis, mechanisms by which they exert their functions, etc. This knowledge will help for early detection of sepsis and its therapeutic targets. The expression of lncRNA nuclear-enriched abundant transcript (NEAT) and TNF- have been implicated in LPS-induced macrophages (Wu et al., 2020).

The choice of feature selection methods for small data sets is a challenging task. While there is luxury for feature selection in big data that would in turn improve model performance, further downsizing of small data poses the challenge of removing important features. While upsampling and downsampling methods are good in data balancing, overfitting is possible due to data duplication. In this study, we applied stability selection to rank our features and used SMOTE to generate synthetic features and balance our data. In this research, we predicted sepsis lncRNAs from all the nine (9) GSE files contained in Zheng's papers, other than from a single cohort.

Materials and methods

Description of Sepsis Expression Microarray Datasets

We used the microarray dataset Zheng's paper (Zheng et al., 2021). In all, we downloaded nine (9) CEL.gz sepsis expression datasets for this study from the Gene Expression Omnibus (GEO) database. Each of these datasets contained their normal controls.

Gene Annotations for Microarray Probesets

We used human annotation files of Ensembl Release 105 (Dec 2021). Affymetrix Human Genome U133 Plus 2.0 Array which has complete transcripts coverage was the only platform used in getting our desired probsets. The probesets with Ensembl gene IDs that were annotated as " long non-coding RNA" in Ensembl were recruited. We got a total of 4995 lncRNA expression profiles based on the aforementioned benchmark from Ensembl.

Selecting lncRNA gene subset for ML Models

We selected only the columns that we were interested in, after which we dropped the rows which had NA in those columns. The GSE95233 data had 54675 samples and 124 probes. After filtering out the NA, we had 54675 probes out of 54675 probes for our GSE95233 data. We observed that some of the probe IDs were mapped to multiple genes, e.g. 1007-s-at was mapped to both DDR1 and MIR4640. We counted the number of probes that were like that and filtered them out. Overall, we had 682

probes that were mapped to multiple genes out of 54675 for our GSE95233 data. We searched for lncRNA genes that corresponded to the probes. Overall we found 1564 probe IDs that corresponded to lncRNA genes out of 54675 for our GSE95233 data. This method was applied on the remaining GSE files. For each dataset of the nine (9) sepsis gene expression dataset, we printed out three (3) output files, namely; gene annotation data (fData.tsv), samples annotation data (pData.tsv) and gene expression data (exprs.tsv) and saved it for ML model training. Table.... shows the number of samples for each dataset accession ID, their probes, samples after filtering NA, number of probes mapped to multiple genes, and number of probe IDs.

Stability Selection Feature Selection

Once applied, stability selection may be novel in feature selection. The whole idea of stability selection is to randomly select a subset of a dataset and rank their features and score them. After running it multiple times, the persistent features will remain higher in the ranking, these features would be given higher score. After Selecting our features through stability selection, we applied Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic features and balance the data.

$x' = x + rand(0,1) * \mid x - x_k \mid$

where rand(0, 1) represents the random number between 0 and 1

Differentially Expressed Genes

Although our interest is not on getting differentially expressed genes, we used limma package and got a list of differentially expressed genes for one of the nine (9) sepsis microarray datasets (GSE95233). Limma package is a tool that works best with microarray datasets (Ritchie et al., 2015). These samples were already normalized by the Robust Multi-array Average (RMA) method. The GSE95233 file can be grouped into control and sepsis patients because they have different expression patterns. We performed differential expression analysis on this dataset using limma package and plotted our PCA, heatmap, volcano plots, and Venn diagram. We started by creating a simple design matrix to compare the control group and sepsis patients.

Principal Component Analysis for GSE95233

To group samples according to their gene expression, we made a PCA/UMAP plots for one of the sepsis microarray dataset (GSE95233) based on the sample characteristics. Fig 1.1: UMAP Plot for gender and survival for GSE95233

Getting upregulated and downregulated genes through Volcano plot for GSE95233

If the ddCt has a negative value, then this means that the gene we are interested in is upregulated, this is simply because the fold change will be greater than 1, but if the ddCt has a positive value, then the gene we are interested in is downregulated and in this case, the fold change will be less than 1 (Livak et al., 2001). The Volcano plot for Control vs Sepsis, Volcano plot for Control vs Non-Survivor comparison, Volcano plot for Non-Survivor vs Survivor comparison, and Volcano plot for Survivor vs Control comparison can be found in the supplementary information section of this paper.

Pre-processing for ML Models

Before setting our hyperparameters for machine learning model training, we took a general look at each of our datasets by plotting the expression distribution. We started by plotting a histogram of total expression in each dataset.

We also made a boxplot depicting each of the nine (9) affymetrix samples. Below is a boxplot for GSE9692 expression dataset.

We can see that the first three (3) datasets from the histogram above are shifted towards higher values in comparison with the rest of the data. If we read the summary of each dataset, we can see that the first three (3) datasets: GSE95233, GSE57065, and GSE28750 are only normalized by RMA method, while the rest of the datasets are also re-normalized by control median. We will re-normalize the first three (3) datasets and then combine all datasets into one big dataset for stability selection and SMOTE before training the models. Also, we added some information to phenotype data, and aggregate expression values per gene symbol. Below is a histogram of our re-normalized samples.

Overall, we had 77 genes and 975 samples after combining our expression values, with class frequencies as follow; Same-Day (H0): 511, blood collection within 24 hours (H24): 48, blood collection within 48 hours H48: 222, Healthy Controls (C): 183, and Post surgical blood collection after 24 hours (PS-H24): 11.

Results and discussion

Machine Learning Models

In all, we had an overall of 1084 genes and 975 samples after we re-normalized all the nine (9) expression files by control median. We then applied stability selection and SMOTE to generate synthetic features and balance the data respectively. We then merged the original features with the synthetic features We separated our dataset into train (0.9) and test sets (0.1) and then used different models to create a predictor. Our test set also came from our re-normalized dataset. We set our random seed value at 42 and then used StandardScaler to scale our features for both datasets.

We trained five (5) simple binary classification models, namely: logistic regression, support vector machine (SVM), decision trees, eXtreme gradient boosting (XGBoost) and random forests. We used five scoring functions to measure the performance of our models, namely; accuracy-metrics, f1-score-metrics, precision-score-metrics, recall-score-metrics and roc-auc-score-metrics.

Model Performance based on our five (5) metrics

We trained five (5) machine learning models, namely; randome forests, XGBoost, SVM, decision trees and logistic regression. We checked the performance of our models on both our training and test sets. Using accuracy as our performance metrics on our test set across our five (5) models, our random forests performed best at 0.97. This is followed by XGBoost at 0.96, SVM = 0.95, decision trees = 0.94 while our logistic regression is the least performance model at 0.93. For f1 score metrics, our random forests still performed best at 0.97, followed by XGboost at 0.96. Our SVM came third at 0.95, decision trees as the fourth at 0.94 while our logistic regression came last at 0.92. For precision score metrics, our random forests performed best at 0.97, XGBoost at 0.96, decision trees at 0.95, SVM at 0.96 while our logistic regression performed least at 0.93. Using recall as our score metrics, our random forests performed best at 0.97, XGBoost at 0.96, SVM at 0.95, decision trees at 0.94, while our logistic regression performed least at 0.92. When we used roc-auc-score-metrics as our score metrics, our random forests

performed best at 0.97, XGBoost at 0.96, SVM at 0.95, decision trees at 0.94, while our logistic regression performed least at 0.92. Best Performing Model: Among all our binary classification models, our random forests performed best at 0.97, 2nd best performing model: XGBoost = 0.96, SVM =0.95, followed by decision trees while our logistic regression model was the least performed model. Feature Importance We succeeded in training binary classification models which captured the differences in groups and achieved high accuracy. We printed the most predictive lncRNAs with the highest and lowest coefficients. We used Gini importance score to extract our important features. Among our models, XGBoost performed best at AUC of 0.97. Based on XGBoost's predictive power, We identified 10 (ten) potential biomarkers for sepsis; namely WDFY3-AS2, NFE4, WDR11-AS1, TP53TG1, ZNF790-AS1, ANKRD10-IT1, LINC01352, CHRM3-AS2, LINC00528, FAM13A-AS1. In line with this research, other researches have shown that our newly found lncRNAs such as ..... sepsis biomarkers (He et al., 2019, Wang et al., 2021). The mechanism of action of these long non-coding RNAs is not fully known. Recent researches has shown that some non-coding functions exerts their functions by forming duplex, triplex and quadruplex structures. Research on how to uncover the mechanisms of action by lncRNAs specific to sepsis will open opportunities for diagnosis and prognosis of sepsis and its therapeutic interventions. With artificial intelligence (machine learning-deep learning), high predictions can be achieved to save cost and energy for in vivo tests and in vitro experiments.

Data availability

Table 1 Contains all the datasets used in this article. All the datasets supporting the conclusions of this article are collected from Gene Expression Omnibus (GEO) repository and public available at their stated accession IDs: GSE95233, GSE57065, GSE28750, GSE8121, GSE9692, GSE13904, GSE26378, GSE4607 and GSE26440 (Table 1).

Supplementary Data

The link to differential expression analysis results and reannotation output files can be found here ….

Abbreviations

- SVM = Support vector machines
- XGBoost = eXtreme gradient boosting
- lncRNA = long non coding RNA
- GEO = Gene Expression Omnibus
- ROC-AUC = Area under the ROC Curve
- UMAP = Uniform Manifold Approximation and Projection
- PCA = Principal Component Analysis

Conflicts of interest

The authors declare no conflict of interests for this article.

Acknowledgments

Funding

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable